DataSpeak – Clusters, Maps and Hotspots: Small Area Analysis in Maternal and Child Health

August 12, 2015

Presentation visuals are available for download (78 pages, 4.3MB PDF).
Corresponding slide numbers are indicated throughout this transcript.

Program Transcript

`Timestamp 00:00 [slide 3]`

Michael Kogan, PhD – Director, Office of Epidemiology and Research, Maternal and Child Health Bureau

Good afternoon and welcome to today's program called **Clusters, Maps and Hotspots: Small Area Analysis in Maternal and Child Health.** My name is Dr. Michael Kogan and I'm the Director of the Office of Epidemiology and Research in the Maternal and Child Health Bureau, which is the sponsor of the DataSpeak series. Those of us at the Bureau and practitioners in the field are increasingly interested in mapping how risk factors and health resources vary between communities. Learning about these factors vary across different counties and neighborhoods can help us design interventions and direct resources to the populations who are most at need. Small area analysis is a tool that can help examine differences in small geographic areas and find variations in health risks, resources and outcomes. Ultimately this information helps inform public health decisions and improves maternal and child health.

Today, we are excited to have three speakers who will take us through some techniques and applications of small area analysis. First**, Dr. Russ Kirby** from the University of Florida will set the stage for us by reviewing basic principles of mapping and small area analysis. Second, **Dr. Michael Kramer** from Emory University will discuss several approaches to the production of statistically robust small area estimates. And finally, **Dr. Tom Stopka** from Tufts University School of Medicine will describe how he used small area analysis methods to identify hotspot clusters of unmet nutrition and public health needs. And now I'd like to turn the program over to our moderator, Sarah Lifsey; Sarah, the floor is all yours.

`Timestamp 01:48 [slides 4 and 5]`

Sarah Lifsey, MPP – Policy Associate, Altarum Institute

Thank you. First I'd like to welcome our presenters and everyone who is in the audience today, thank you for joining us. Before we begin our presentations, I just have some brief technical guidance for everyone. First I'd like to call your attention to the DataSpeak website, which we hope you'll visit after today's program. On the website you'll find archive of all the DataSpeak programs going back to 2000 and the slide on your screen shows some of the most recent programs that are available and the address that you can use to access them.

I'd also like to point out that you are able to download today's PowerPoint presentation directly from the screen that you are seeing right now. Once you are finished, you may click the continue button to proceed with the presentation. At the completion of the program, we'll be having a question and answer session. So now I would like to turn to our first speaker, Dr. Russell Kirby. Dr. Kirby?

```
Timestamp 02:39 [slides 6 to 30]
```

Russell S. Kirby, PhD, MS, FACE – Distinguished University Professor and Marrell Endowed Chair, Department of Community and Family Health, University of South Florida

Yes, I'm ready to start. I'm going to actually go to the next slide; but my task is to give you a brief review of some of the basic principles of mapping and some of its uses for looking at health phenomena from the spatial perspective and introduce the approaches of small area analysis. And I'm going to do that by way of a couple of examples. Before I get to that though, I wanted to make a couple of points.

One is that maps, even though everybody likes to think of them as distinct kinds of entities, they really are just a very specialized form of data graphics and so it's useful to think about basic principles of visual display of data also when we think about maps. So I'm giving you here some references that might be of interest. Many of you are probably familiar with the work of Edward Tufte; I've referenced his first book, but he actually has several others that you might have seen. If you're interested in getting more into the theory of data graphics, Bertin wrote a classic book on *The Semiology of Graphics* published in 1967; there was an English translation in the 1980s. And then Wilkinson, who many of you may know as the developer of SYSTAT, one of the statistics packages that was in vogue in the '90s, put together a *Grammar of Graphics* that can be operationalized using… that allows you to do a wide array of things.

But anyway, I also wanted to emphasize that when you think about displaying data spatially, frequently we think of the map as the end result of our analysis. And I'd like to encourage everyone on the phone to think about maps more as the starting point for analysis than the ending point because when we think about a spatial pattern on a map, what we're really interested in are what are the factors that drive that particular pattern? In other words, what are the spatial processes that are involved?

And so what I'm going to do in terms of examples that illustrate some of the aspects of mapping are to look at, um, two different examples; one looking at Hispanic population in the US and the second looking at birth data for US States. And all of the maps I'm going to be showing are examples of what we call choropleth maps and a choropleth map is one in which the map is divided into small areas and then each of the areas is categorized based on where it's value falls within a range of values.

And so, firstly, the first map I have is from the Census Bureau and it shows the distribution of the Hispanic population in the United States. And this is a very interesting map; it shows areas that have high density Hispanic populations, and you can see that the majority of counties in the United States in the year 2000 did not have very many. But take note of the categories, you know, the highest is 50% or more, the lowest is less than 6% and the next highest is 6 to 12.4%. When we look at the next map, this next map is also published by the Census Bureau, actually in the same publication as the first one, but for 10 years later.

And you can see if you flip back between these two, there's what seems to be a larger area of the United States that's now falling into the higher categories for the distribution of Hispanic population and there certainly are many counties that were in the lowest category that have now moved up. But, when they made this map, the categories are not the same and that's one of the basic principles of mapping; when you're showing phenomena across time is that if you want the reader to be able to really visualize what's going on, you want to make it as easy as possible for them, which is something the Census Bureau did not do in those maps.

Here's another map; this is a map that shows the race and Hispanic diversity for the United States. And I'm going to go to the next slide just for a moment. This is using the diversity index, which is calculated as the probability that two persons selected at random from a counties population would either be of different races, or that only one of the two would be Hispanic. And so when we look at this map, what we find is that when we think not just about Hispanics but the broader range of race ethnic groups in the United States, there's actually quite a lot of diversity. It tends to be focused in the American South and Southwest, but also in major metropolitan areas and also in the Western United States, in areas that typically are areas where there are reservations. So that's another way to look at it.

But another thing to think about when we look at the distribution Hispanic is that not all Hispanics are alike. And this map shows the basic pattern of origin of individuals who are Hispanic in the United States and what you can see is that in many states, almost everyone is Mexican origin. However, there are areas in the Northeastern United States where other origins are more common, in particular, Puerto Rican that we have in the New York metropolitan area and also in Massachusetts and Connecticut. And then if you look at Florida, you can see that there are where Cubans are the most common. So there is quite a bit of diversity that you see within this particular group.

And then this map… this is not a map, this is a data graphic which shows that there's a lot going on beneath the surface, in terms of the distribution of the Hispanic population. And this is a population pyramid showing for 1900, 1950 and 2000. And the pyramid in 1900 was similar to what we typically expect for a population pyramid that most individuals at the youngest ages and then tapering off as we get higher. But that's changed markedly in recent years and the pattern that we see for the year 2000 has a bulge right in middle adult ages, you know about 35 to 45 to 50 or thereabouts. And that tells us that there are going to be some interesting changes in this population in the future because we're looking to have potentially more individuals who are Hispanic aging out of reproductive years.

Now, it is possible to lie with maps, and in fact all you have to do is read any map in a newspaper to learn how to do it, because that's where we get all our best examples from. But Mark Monmonier wrote a book about 20 years ago called *How to Lie With Maps,* it was actually a second edition, and almost all of his examples are MCH. So I would definitely encourage you to look at that book if you haven't seen it and I wanted to show a series of maps that he put into an article that came out about 10 years ago or so. It’s in a Journal called *Statistical Science.* And what this is showing is a series of maps all showing the crude birth rate for the United States by state.

And what you can see in this first map is that he’s got categories that break out particularly the states that are in the highest range for the crude birth rate and Utah being the highest and states like Texas and New Mexico being in the second group and so on. And most of the states in this map are in the third group, which is 13.1 to 15.2. Another way you might have made this map would have been to classify all the states based on replacement population, which is roughly 2.05 or 20.5 per 1,000 it would be. And if you did that, you’d find only Utah would be in the highest category.

Here’s another map where he’s broadened the range for the higher strata, now there are all sorts of states that are lighting up. And if you looked at this map you might conclude that pretty much the whole Southwest and much of the American south has the highest crude birth rates. But looking at the previous map, we know that’s not necessarily true. In the third map, he’s using the lowest category, you know with a wider interval and now you’re seeing that 3 quarters of the states are falling into the lower category. Again, it’s all a matter of what your intent is as to how you want to classify your data.

This next one is really emphasizing the states, making it look like many of the states have the high rates by using 13.0 to 13.9 as the second highest category and 14.0 and higher as the highest category. Again, it all depends, you know, what your point is with how you want to display the data. And then this map is showing not the crude birth rate, but the actual numbers of births in each state. And here what’s been done is to use 120,000 plus as the cut off for the highest category. And by doing that, it causes a number of states to show up in the highest category, which may again potentially be misleading.

And then finally, this one shows using dots representing 10,000 births; gives you an idea of bigger means more. So these are all ways that we can display data differently but using the same underlying data. And the whole point is that any data series that you might have, you can make almost an infinite number of maps and your challenge is to come up with, what’s the most important pattern within the data that you want to display? And when Michael starts his presentation in a few minutes, he’s going to pick up on this particular theme.

So, getting into the whole issue of why do we want to do this in the first place; well, back in Epidemiology 101, you learned that we focus on person, place and time and place is very often forgotten, other than a way to collect the data. And so small area analysis provides us with a set of methods that allow us to analyze phenomena at the local level at different scales, counties, zip codes, census tract and potentially even latitude and longitude, if we have access to that level of data.

And I don’t need to convince you, I hope, but there are many phenomena that cluster in space and locations or regions that have spatial gradients that vary across state or region or vary dramatically by country. And examples of clusters include homicide, injury, environmental exposures and their outcomes, occupational illnesses, food borne illnesses and so on. Here’s a map that was published by our colleagues from New York State Department of Health a few years ago where they did a cluster analysis using a modified version of SaTScan to identify clusters of low birthweight in the state of New

York and it does. You know, there are areas that you can identify that have statistically elevated rates of low birthweight.

Spatial gradients, there are many examples, but classic are multiple sclerosis, which tends to increase in prevalence through the mid-latitudes, both north and south of the equator. Environmental contaminant exposures— a classic example is a number of studies have been done showing patterns of birth defects associated with radioactive fallout from the Chernobyl incident in the 1980s. And then there are many examples of regional and national variation, you know asking questions like why does the US have higher infant mortality rates than western nations? Or, why do states in the US south tend to have higher rates of C-section and things of that nature.

And we also think about places making a difference in more complex ways such as what we tend to call the immigrant paradox, that women who migrate to the United States from a particular receiving country or US is the receiving country, from a nation of origin tend to have better outcomes than members of their own national origin group in the country where they were born. Or, what is it about socioeconomic status and education? You know, why do persons of similar educational attainment who live in poorer neighborhoods tend to have poorer health status than those with that same level of education living in a higher income neighborhood? And these are all things that small area analysis can help us to start to understand.

So, in summary, an infinite number of maps can be created for any variable that we want to measure across small areas. Maps allow the user to better understand patterns in their data, but the ultimate goal should be to understand the processes creating the patterns and then small area analysis is a set of tools that enables us to better understand patterns in health data that we can display spatially. And I'm now going to turn over for the remainder of the webinar where you'll get some examples of applying small area analysis in maternal and child health.

`Timestamp 17:22`

Sarah Lifsey, MPP

Great. Thank you so much Dr. Kirby. As a reminder, if you have a question for our speakers, you can just submit it online at any time using the questions form at the bottom of your screen. So next I'd like to turn to our next presenter, Dr. Michael Kramer. Dr. Kramer?

`Timestamp: 17:43 [slides 31 to 46]`

Michael Kramer, PhD, MMSc – Assistant Professor of Epidemiology, Rollins School of Public Health, Emory University

Thank you, Sarah. Thank you, Dr. Kirby. All right, so as Dr. Kirby said, I'm going to sort of take those general ideas and move on to sort a next set of challenges we have in small area estimation and briefly discuss the problem that faces any analyst who's trying to say something about data for small areas where there might be either few events or a small population at risk. As Dr. Kirby has already reviewed, we might be interested in these small areas because of the way that places encode prophesies which affect health. For instance, we're… we may not be interested solely in the overall average prevalence of an outcome, but we're also interested in sub-population heterogeneity; in other words, we might be interested in how does one group fare as compared to another, or perhaps how does one place fare as compared to another.

For those of you who may be in state and local health departments, you may have been asked by community or programmatic or legislative stakeholders for spatial representation from maps of the health of women and children. And to produce those, you need small area estimates. Finally, another reason we might need small area data are as inputs to subsequent analysis. It may be that we're interested in doing a cluster analysis or spatial regression or some additional step of data analysis and the input we need is a stable rate estimate.

Producing these small area estimates has several pitfalls though and there are problems that we have to wrestle with to do this. One concern is just that of privacy; if you are a data manager, you work with data, you're aware of the fact that we may be very concerned about inadvertent disclosure of protected information in the form of small area maps. And so one thing people will do to deal with this are things like aggregating across geographic areas or pooling across year, sometimes data suppression or methods of geomasking.

But a second problem, one that we're going to focus on a little more today is that of the statistical stability of any risk or rate or proportion in the instance where there may be few events in the numerator or small populations at risk in the denominator. In MCH, we see a variety of approaches to address this kind of problem. For example, MCH and CDC data may occasionally be suppressed when counts are small.

Another approach in design-based surveillance systems, such as PRAMS, for example, where you might take a stratified sample that oversamples important subgroups, say very low birthweight infants in order just to have an adequate sample size. And then uses survey design weight to represent the whole population. But the option that we're going to talk about primarily today is this third one and that is model-based estimation.

As an example today, I'm going to be talking about a relatively rare event, so something that could have sparse events, and that's the birth prevalence of very low birthweight infants or those who were born less than 1,500 grams. So here's a map of the prevalence among non-Hispanic white women in the lower 48 states in 2005 to 2007. And you can look at this and see some spatial patterns that are apparent. Just for information, hopefully you can see the size of the… you can see the legend, read it okay on the screen. The map symbolizes quintiles of states according to very low birthweight prevalence and looking at the legend, you can see that the lowest risk state is about 0.84% and the highest risk state is about 1.47%.

But, we might be interested in how much the risk varies not only between states, but also within states. In other words, we might be interested in how much heterogeneity there is across counties. So in this county level map the first thing that is notable is that in almost every state there is variation such that most states have both a very high and a very low risk county.

So something that was not apparent from the state level map. If you look again at the legend in the lower left, now you see a slight difference and that is that the lowest risk or prevalence county is 0% and the highest is 23.08%. In other words, there’s a pretty dramatically different picture in the range of rates than we got for the state level map.

So this example illustrates two problems that are really common from the direct calculation of risk from raw, observed data for small areas. So there’s the first problem, which is one of extreme values. When the population at risk that we’re looking at is very small, the presence or absence, the addition or subtraction of even a single event can dramatically change the ultimate prevalence or risk estimate.

So for example in that county map that I showed a second ago, there were actually 273 counties that had an estimated prevalence of 0% very low birthweight. Well of course it’s perfectly reasonable that there are counties that had 0 very low birthweight births during the 3-year period, but from a public health perspective, we’re really interested in ascertaining the true underlying risk in the county, and it’s unlikely that there’s any population that has a true risk of 0%.

On the flip side, there were about 25 counties had rates that were 3 to 4 times the national average. So again, extremely high prevalence of very low birthweight is totally possible, completely possible but since most of those counties actually had 50 or fewer births in 3 years, we should certainly be cautious of making very strong inference about those rates. So that’s the first problem, extreme values.

The second is related, and that has to do with the certainty, the statistical precision of the estimates themselves. So we can imagine that to make that map I showed a moment ago, each county we have an estimate the prevalence of very low birthweight, but we can also calculate a standard error, and one way to combine these is this measure of the relative standard error, which is the standard error divided by the estimate itself. And some people have suggested that a relative standard error greater than 30% is… gives reason for caution or concern about interpreting that value. In other words, it may be a statistically unstable estimate. And applying that rule in that map of US counties, more than half of the counties have an unstable estimate, a relative standard error greater than 30%.

So an approach to these problems… there are several approaches. One approach is to try to borrow some sort of auxiliary statistical information. We borrow that information in order to end up with more stable and trustworthy estimates of each small area. So the raw data approach presumes that all we know for each county is the numerator and denominator and it treats each county as an isolated island without any other context when, in fact, it’s not the case. What we know is the entire distribution of rates. We know the distribution, how… what the minimum, the average and the maximum is across the country.

So our approach today is to talk about methods that estimate a value for each county that uses the information from that county, but when necessary, also borrows information from a known distribution, the distribution of rates across some other population. This approach is called parameter shrinkage, because what we’re doing is we’re shrinking the previously uncertain value toward some local or regional or national average. And let me just briefly show you 3 examples of how we can apply this method.

This first one is here that I've called Aspatial Multilevel Regression, tis one actually requires no GIS or spatial software at all. It just requires pretty routine statistical software that you might have available like SAS or STATA or R. The approach is basically simply a multilevel or random effects regression where we nest counties within states. So if there's a county with an extreme value of very low birthweight, extremely high or low, and it has a large number of events, the model will not alter its prevalence. But if we had an extreme value county with very few events, small amount of data, that county will be shrunk towards the average of the state. So that's the shrinkage.

Now the second method, called Aspatial Empirical Bayes, is related. In this case though, we're shrinking each county towards the distribution of all the counties in the US, not just towards a single state. So we're using a different normal distribution or different prior information. And the last one there, finally we can use an explicitly Spatial Empirical Bayes. And by spatial, what I mean here is that we're incorporating some knowledge about which counties are neighbors to which. And then we use the local average of very low birthweight among the neighbors as the expected value, which will be combined in some way with the index county estimates. Okay, so let's get to some maps and think about what this means.

First, this is the raw data—this is without any shrinkage or anything. The way I've symbolized this map is important to talk about. As Dr. Kirby was saying, there are many ways you can symbolize it. This is called a box plot symbology and I've chosen it because I'm trying to emphasize outliers. So just look at the box plot on the right-hand side of the slide here. You can see that most of the counties in the box part of the box plot are in a narrow range of very low birthweight risk around 1%. And those counties are colored light blue. But there are also counties that are in the whiskers of the plot—the line part—and especially on the high end, there are several extreme outliers that are colored red… dark red on the map.

So if you look at the map itself we can see a swath of counties running north south from the Great Plains down into west Texas where there seem to be examples of both extreme highs and extreme lows. But knowing a little bit about the population counts in those counties, many of them are rural counties with small population. Also note in the lower left-hand corner I've just identified the number of counties that have an unstable or an extremely large standard error, in this case, more than half of the counties.

Now the first shrinkage map that I want to demonstrate is the multilevel model, the aspatial multilevel. And the… looking at the box plots, the first thing you notice is that this approach dramatically reduces extreme outliers. The entire distribution is substantially compressed compared to the previous map there; this is the raw data and this is the aspatial multilevel.

When we look at the map itself, we can see that there are a few "extremes." Again the extremes in this case are much closer to the mean than they were in the previous, but a few extremes sort of scattered around the US. Looking in the lower left-hand side of the slide, you can see the number of counties that had an unstable rate in this case, and it's only 26 instead of 1,700 counties. So this reduced the outliers and it also substantially stabilized the precision of each rate estimate.

The second method is the Aspatial Empirical Bayes and it differs in a couple of ways, but importantly in that it doesn't specify the states as the "norm" for each county, instead allowing information from the entire US to inform each local estimate. Again, looking at the box plot on the right, the range of estimates is much narrower than the raw data, but it is a little bit wider than the multilevel model, which is not necessarily bad.

And looking at the map itself, we can see extremes, so the darker colors are scattered along the Appalachian region and across some parts of the south, and both of those are areas where we might expect to see such extremes. Also just as with the multilevel model, the standard errors for each estimate are much smaller than in the raw data itself.

And the third approach that I'm showing here for comparison is the Spatial Empirical Bayes where instead of shrinking the local estimates to one overall national distribution, we're shrinking to a local distribution, specifically to the average of all of the adjacent counties. So here the box plot is not as extreme as in the raw data, but you can see that there are more outliers than in either of the aspatial shrinkage methods.

Looking at the map itself, we can also see the emergence of some patterns of clusters of both highs and lows in counties across the US and in some areas that we might expect them to be. The relative standard error is high in a few more counties than the aspatial method, but it's still substantially better than the raw data.

So our goals, to just state them again, were twofold; one is to reduce unrealistic outliers. In other words, things that are not believable given the data either at the low or the high end, but also trying to preserve the true underlying variation that drove our interest in small area analysis in the first place. In other words, we don't want to smooth it so much that everything goes away.

So on the first count, the two aspatial methods dramatically shrink outliers and all 3 methods dramatically improve the standard error for estimates compared to just the raw data calculation. When it comes to distinguishing between extremes, unusual highs and lows, the Spatial Empirical Bayes may have maximized this approach. It maximizes differentiation of highs and lows.

The choice of which method works for a particular situation depends really on the question and the balance between these two goals I have in this slide—these two things are kind of competing with one another. But I should also point out that I symbolized these maps to accentuate outliers and not to look at the distribution in the bulk of the counties, and that when we look, for example, at a quintile map, all three methods identified meaningful and important heterogeneity.

Now I've talked about just 3 related methods, but there are really dozens of ways to produce small area estimates and I certainly don't have time to talk about them, but I just wanted to mention these broad classes that may be interesting to some people. For example, fully Bayesian disease mapping is another approach to small area estimation. Kernal Density smoothing, spatial kriging and this wonderfully named iterative weighted head-banging algorithm, which just sounds like something you want to do.

Okay, so how might you actually do any of this? As I already touched on, the aspatial multilevel approach is something that you can do if you have access to pretty routinely available software, either commercial software like SAS of STATA or SDS or open source software like R. From the point of view of the Empirical Bayes Estimates, there are a number of ways to do it, but I want to really point out there's one software package that I like because it is freely available and relatively easy to use, and that is the GeoDa package for exploratory spatial data analysis. It can do a lot of things, but one of them is it can quite readily and easily estimate Empirical Bayes Spatial and Aspatial small area estimates.

So, in conclusion, our goals were really to maximize the information that's available in our data so that we can describe small area variation in health outcomes. But we want to do so in a way which is both interpretable and statistically robust. I've discussed several methods of model based estimation and the way they work is by borrowing statistical information from neighbors or from other areas in order to stabilize local values, hopefully to minimize unrealistic outliers and ideally to maintain interesting and the relevant small area variation that, as I said, brings us to this method to begin with.

I've included a set of a very small portion of many possible references, but these are some texts and some papers that are particularly useful detail about the statistical background and the operationalization of these methods. And I'll turn it back over to you, Sarah.

```
Timestamp 34:14
```

Sarah Lifsey, MPP

Okay great. Thank you so much, Dr. Kramer. So next, I'd like to turn to our last speaker is Dr. Tom Stopka. Dr. Stopka?

```
Timestamp 34:24 [slides 48 to 75]
```

Thomas J. Stopka, PhD, MHS – Assistant Professor, Department of Public Health and Community Medicine, Tufts University School of Medicine

Thanks very much, Sarah. So I'm going to try to talk to you about an applied example of small area analysis as we used it in California; so good afternoon to folks in the east and central US and good morning to our friends in California. I'm going to apply this example in talking about WIC. And I'd like first to acknowledge some of my colleagues on this work that we conducted a couple of years ago at the California Department of Public Health, Drs. Krawczyk and Curtis; the California WIC Program, Dr. Pat Gradziel and one of my mentors in GIS and spatial analysis, Dr. Estie Geraghty. If you'd like to learn more about this study and about the methods that we applied, please take a look at our paper in the *American Journal of Public Health.*

So WIC, for folks that may be less familiar with WIC, is the special supplemental nutrition program for women, infants and children. It reaches approximately 1 in 4 pregnant women and roughly 50% of all infants born in the US who participate in WIC and more than half of pregnant women enroll in WIC during the first trimester. Now in California, WIC agencies provide services across the state in local sites to approximately 1.5 million women, infants and children through more than 600 sites.

And some of our questions were where are statistically-significant, WIC-eligible women located within California? And where do micro-level clusters exist? And that was to basically answer a question when there were some challenging times in terms of budgeting in California. And I want to emphasize that GIS and spatial analysis can be used in good times and in bad. In 2012, the WIC program faced a potential cut across the nation of over $800 million dollars and that could have impacted up to 500,000 low income women and children who would have been denied services. We thought that in California it might be useful to conduct some spatial analyses and small area analyses to assess needs and unmet needs across the state.

And when we decided to look at clustering in California, we came up with a definition that we really wanted to explore, and this is for WIC-eligible women, at the bottom of your screen. And these were women who receive Medi-Cal during pregnancy, that is to say they were WIC-eligible, but they were not actually receiving WIC services. So we conducted a multistep approach to merge 2 large data sets; birth statistical master file; there are about a half million births in California each year. We merged these data with the WIC-ISIS data, so this is the data systems for all WIC sites in California.

Our outcome of interest ultimately was WIC-eligible women not receiving WIC services, and there were about 30,000 women who fell into this category. And we were specifically looking at the density of WIC-eligibles per square mile on the census tract level. And during this timeframe there were about 7,000 census tracts across California.

We first started to portray the data descriptively through schematic GIS maps and choropleth maps and some dot density maps and then started to use some spatial epidemiologic approaches and spatial analyses that were a bit more complex in ArcGIS 10.1. And I just want to emphasize there are a number of different spatial analytical methods that are out there, and Drs. Kramer and Kirby referred to several. And just to emphasize, there's no one approach, no one map, no one graphic that can tell the whole story—it's really the compilation and the complement of different approaches that can really start to get to the whole story.

So initially we looked at the macro-level, we looked at the big picture. These are the 58 counties in California and we were interested in the number of WIC-eligible women who were not participating in WIC across counties. And that started to tell us a little bit of a story about which counties perhaps had a higher burden of unmet need. We started to dig down a bit deeper to a smaller-level analysis and here we are looking at medical service study areas; there were 541 in California. Again we're looking at the number of WIC-eligible non-participants and this starts to tell us a little more detailed of a story in terms of some small areas that perhaps have a higher burden of unmet need.

And then we turned to some dot density maps that helped us to get down perhaps to the census tract level and a little bit below, to better understand perhaps locations in the central valley and in the bay area and in Southern California that might merit more attention in terms of addressing unmet need as it relates to WIC services. But still, these maps were largely descriptive and the outline questions for us were how do we know that these patterns are not due to chance alone? And where are there statistically-significant clusters of WIC-eligible women located in California?

So we needed to conduct Hot-Spot analyses to do that. So Hot-Spot analysis, or Getis-Ord Hot-Spot analysis, which produces a GI* statistic is a tool that's available in ArcGIS and we used it to pinpoint locations of clusters. And in doing so, we're looking at each feature, in this case census tracts in California within the context of neighboring census tracts. And the census tract with a high value was found to be statistically-significant if… a significant hotspot if it was also surrounded by other features with high values.

The local mean for a feature or a census tract and its neighbors is compared proportionally to the global mean for all census tracts in California. And when the observed local mean was much different than the expected local mean, than that difference is too large to be the result of random chance; we found a statistically-significant Z-score result.

So, in order to take the first steps at finding clusters, we had to first determine what the exemplar sphere of spatial influence was. What is the space or the distance frame at which clustering is most important? And in this map, I'm showing you a number of census tracts, and the green dot in the middle of that census tract in the middle of your screen is the Geo-centrite or the geographic center of that census tract. And we have to determine whether looking at local neighbors at 1 km, 5 km, 15 km, or 25 km is most appropriate in determining the clustering of WIC unmet need.

And to do that, we use a 5-step Geoprocessing approach. In the first step, calculate the area for the polygons—the census tracts—and we excluded the areas that were 1.5… greater than 1.5 standard deviations above the mean square mile area for census tracts to account for variation in polygon size and, at least for a bit, would leave out the outliers, in terms of the large census tracts.

We then needed to find the appropriate spatial scale for selected tracts, and we looked at the distance from each tract to its 2 nearest neighbors. And this provided us with 2 parameters that ultimately we could use in our next steps for our starting distance and our incremental distance in our spatial auto-correlation analyses.

So in step 3, we conducted incremental spatial auto-correlation which produces a Moran's I, a Z-score and a p-value, and we were able to determine the multiple distances outwards clustering peaks and to find the distance of statistically-significant peaks. And so, in our analyses we found, for instance in this example, that the peak was at about 5,000 meters; the distance was 5,000 meters where clustering or spatial auto-correlation was most intense with the highest Z-score.

So we used that distance in our subsequent steps; first in our creating a spatial weights matrix in order to look for the connectivity and the weighted connections of the different census tracts within California, and then ultimately to conduct the Hot-Spot cluster analysis. And when we produced our results for hotspots, we were able to pinpoint statistically-significant clusters initially on the state level and that drew our attention to specific regions of the state and specific counties. And then we were able to rerun our analyses to focus on specific census tracts within selected counties.

And our results in terms of the Z-scores presented, indicated where there were larger Z-scores, which meant there was more intense clustering of high values, meaning to say it was a hotspot. And where there were smaller Z-scores, there was more intense clustering of low values, or low WIC unmet need or a cold spot.

So I'll share with you some of our results on the statewide level initially and this map portrays shows us statistically-significant census tracts in terms of higher densities of WIC-eligible women who weren't receiving services; now those are hotspots indicated by red. And the blue represent cold spots or statistically-significant clusters of census tracts that had lower unmet need with regard to WIC services.

So we then conducted sub-area analyses to get down to the micro-level analyses within these 4 counties: Sacramento, San Francisco, LA, and Fresno. And in San Francisco, we were now able to highlight statistically-significant hotspot clusters for WIC unmet need in specific neighborhoods in San Francisco, like the Tenderloin District and the Mission District. And overlaid on these hotspots are the green stars which represent the WIC centers that were existing at the time. And so we can juxtapose these two different layers of data to try to determine perhaps where additional services might be needed or where some services are less needed.

We did the same thing for LA and found that the central LA, the Long Beach area, and northwest area of LA County had hotspots for WIC unmet need. And then in Sacramento, we found that there were clusters of WIC unmet need in the north central area and in the west of Sacramento. And in using these data and these results, as well as other data and analyses that were available on the local level in Sacramento, one WIC site decided to ultimately open a new WIC center and within the first 3 months of this WIC center opening, there were several thousand clients enrolled in this center, so it appears that folks made a good decision in placing their WIC center based on their local data and some of the analyses that we did.

So in closing, I'd like to emphasize that we used this 5-step geoprocessing approach to be able to determine hotspot analyses using a systematic, rigorous, and objective approach. We were first able to detect state-level hotspots to locate statistically-significant clusters of WIC-eligible women in California counties and then we reran the analyses on the local level to determine local neighborhood hotspots for WIC unmet need.

And ultimately these findings as well as other findings and other maps and spatial analyses we conduced could help to inform WIC program in funding decisions at the state and local level. And these approaches, as well as a slew of others could be used in other states, in other health departments, and in other counties across the US.

Here are a few of the references that we used that you might be interested in looking at. And I'd be happy to answer anybody's questions, either online or after the talk. Thanks for your time.

```
Timestamp 47:38 [slide 76]
```

Sarah Lifsey, MPP

Great. Thank you so much, Dr. Stopka and thanks again to everybody who has presented today. It's been a great program, and we already have some questions coming in. I will start with a few of the online questions that have come in.

Questions and Answers

Sarah Lifsey, MPP

And the first question I have is from Donna, and it's for Dr. Kirby in reference to the multiple birth rate maps he was showing and discussing.

Of those maps, which one is true, and how do you decide which of the maps is true?

Russell S. Kirby, PhD, MS, FACE

Okay, yeah, that's a great question which doesn't have a direct answer. The long and short is that all of the maps that I showed are true; they all reflect where each of the counties falls within the categories that were chosen for that particular map. The problem is that depending on what your goal is in terms of presenting the data, you might want to use different sorts of maps.

My personal preference would be to do something very similar to what Dr. Kramer did and really have a focus on understanding outliers, and those are often the most interesting, because they frequently drive the overall pattern, and they also sometimes are outliers because of error, and by identifying them, you can figure out what the sources of error might be rather than focusing more on broad categories for central tendencies. But it all depends on what you're trying to do, and if you're doing environmental exposure research, you might prefer to focus on classifying your localities by quintiles or quartiles, and put them in the different categories in that way. But it really depends, and I don't know if Dr. Kramer wants to add anything to that?

Michael Kramer, PhD, MMSc

Yeah, I agree with what Dr. Kirby said; I find that I often use quintiles for general maps that I'm showing to other people when I'm trying to just show the full distribution. But, in my process of data exploration, I choose a map based on what I'm trying to get at, so that notion of outliers versus set cut points. Another consideration is for the thing that you're mapping, is there a natural and actionable cut point? For example, it may be that you're trying to identify the counties that met a Healthy People 2020 goal; and therefore, the goal defines a cut point. So there could be policy or programmatic or evaluation reasons why you'd choose something, and as long as you're transparent about what your choice was based on, I think that's a fair representation of data. The concern is when you use your way of visualizing the map as a way to obscure the data rather than to transparently illuminate it.

Thomas J. Stopka, PhD, MHS

Yeah, and I agree with both speakers; I tend to use quintiles myself. As epidemiologists we also like rates, so it's important recognize that in portraying counts, perhaps you're getting at burden of disease or burden of low birth weight. But in portraying rates, you're normalizing the data across a population, which is another potential consideration.

`Timestamp 51:06`

Sarah Lifsey, MPP

Great, thank you. Our next question is from Debbie, and it is:

What guidance would you give for mapping rare conditions which have very small numbers and where maps might reveal the location and identities of affected individuals?

Michael Kramer, PhD, MMSc

This is Dr. Kramer. I can weigh in on that and that is, somewhat… I think the answer depends; it depends on how identifiable that can be. There is most certainly ethical concern about the representation of surveillance or any kind of health data on a map, and if your map visually makes a person or a small group of people pretty identifiable, then that's no different from having the sparse cells in a small survey in a table in your paper. So I think there's the ethical concern about how to manage that, and there are a lot of things that go into that including probably the preferences for the data holders who may have guidelines.

One thing that mapping can do is, it can try to aggregate enough events together to still observe spatial variation without identifying individuals. But, you know, in the case of birth defects, there's probably a floor about how small you can make your maps before they become identifiable. That's for dissemination though; it could be that if you're in a state health department where you have access to that small area analysis information, you use that in some form of analysis and then the way it's disseminated to the public it is still preserving the privacy of individuals.

Russell S. Kirby, PhD, MS, FACE

Yeah, and if I could also jump in on that, and this does come up a lot with birth defects, but it also comes up with other conditions that we're interested in in maternal and child health. One of the things that I'd like to point out though, is that if you use some of the methods that Michael presented and also mentioned for smoothing and kriging and other kinds of techniques, they enable you to use the information about the rare events in small areas but not actually… that's not actually what you end up displaying in the map. Because the information is smoothed across areas and so if you use those kinds of methods and you're allowed to have all the data to start out with, you can sort of finesse the issue of having very small numbers in particular locations by using these methods which, you know, generally fall into kind of a borrowed strength kind of a framework.

```
Timestamp 53:58
```

Sarah Lifsey, MPP

Great. The next question I have is from the Vermont Department of Health:

> **How many analysis areas are required for these various small area analysis techniques?**
>
> *Thomas J. Stopka, PhD, MHS*
>
> This is Dr. Stopka. I'll chime in and say, the more you have the better, you know, if it's possible to look at small area analysis down to the census tract level or the block group level that can be particularly helpful. There's no right or wrong answer; however, you'll have more statistical power and less susceptibility to chance dictating your results the larger your sample size gets. So the larger your number of… analysis that are included.
>
> *Michael Kramer, PhD, MMSc*
>
> This is Dr. Kramer. If you think about it as a sample-size issue of individuals, it's the same idea. If we're analyzing spatial areas, you're unit of analysis is the spatial area, and so if you have 4 counties, you have an *N* of 4 sample and so you could infer the amount of power you would have if had 4 individual people in your study or in an analysis. There's obviously a limited amount you could say about clusters with an *N* of 4.
>
> *Russell S. Kirby, PhD, MS, FACE*
>
> But I'd like to point out another thing which I think was alluded to in some of the presentations, and say you're looking at low birthweight as a phenomenon, and you have a spatial unit where there are 7 low birthweight cases out of 100 births and you have another where there are 70 low birthweight cases out of 1,000 and yet another where there are 700 out of 10,000. They all give you the same value of 7% or 7 per 1,000. But, if you analyze those without incorporating some measure of the standard error, you will potentially end up with misleading results because the 700 per 10,000 is a much more precise measure than the 7 out of 100. So it's not just, you know, how many areas you have, but also the relationship between the number of events and the number of outcomes.

```
Timestamp 56:14 [slides 77 and 78]
```

Sarah Lifsey, MPP

Great, thanks so much. Well I'm afraid that that is all the time that we have for discussion today. I know we do have some questions we didn't get to, and answers to those questions that we weren't able to address during this Q&A period will be posted in writing along with the program archive. And you'll receive a link to the program archive when it's posted. That archive will be available on the DataSpeak website in the next few weeks. You can access it at your convenience. And, if you think of any more questions, you can submit those to us via email through the end of this week using the email address dataspeak@altarum.org.

So, before you go, we would like you to know we will be broadcasting more DataSpeak programs in the coming months, and announcements about these future programs will be sent out via email to everyone who registered for today's program, as well as on the DataSpeak website.

And finally, before you logout, we'd really appreciate you taking a moment to provide us with feedback on today's program. It's really important to us that we have your input on this session as well as your recommendations for future programs. So to fill out this very short survey, simply click on the evaluation link on your screen now, and a survey will open up in a new window.

Lastly, I would like to thank our 3 speakers for providing us with this great program today. Today's program is now complete; thank you so much for joining us, and I hope you have a great afternoon.

```
Total time 57:37
```

About DataSpeak

The Maternal and Child Health Bureau's DataSpeak webinar series is dedicated to the goal of helping MCH practitioners on the Federal, State, and local levels to improve their capacity to gather, analyze, and use data for planning and policymaking. DataSpeak is funded by the Maternal and Child Health Bureau's Office of Epidemiology and Research under the supervision of Gopal Singh, PhD.

August 28, 2015